

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-175305

(43)Date of publication of application : 21.06.2002

(51)Int.Cl.

G06F 17/30

C12M 1/00

C12N 15/09

C12Q 1/68

(21)Application number : 2000-370947

(71)Applicant : BIOMOLECULAR ENGINEERING RESEARCH INSTITUTE

(22)Date of filing : 06.12.2000

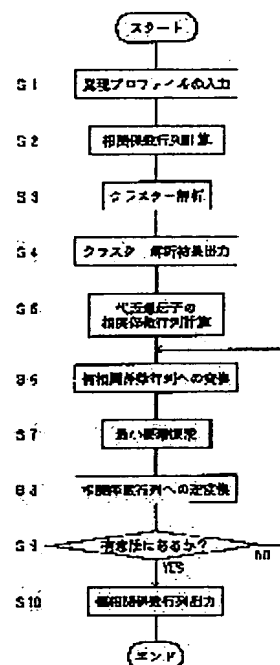
(72)Inventor : FUJI HIROYUKI
HORIMOTO KATSUHISA

(54) GRAPHICAL MODELING METHOD AND DEVICE FOR INFERRING GENE NETWORK

(57)Abstract:

PROBLEM TO BE SOLVED: To provide graphical modeling method and device, capable of quickly and precisely inferring a gene network from a manifestation profile.

SOLUTION: The method includes steps in which a gene manifestation profile is inputted, correlation coefficient matrix for genes is calculated, cluster analysis is executed, a cluster analysis result is outputted, and the correlation coefficient matrix of representative genes is calculated. Then transformation into partial correlation coefficient matrix is executed, minimum elements are retrieved, inverse transformation into the correlation coefficient matrix is executed, and significance is determined. When there is no significance, as a result of significance judgment, operation is returned to the transformation into the partial correlation coefficient matrix, but when there is significance, a partial correlation coefficient matrix is outputted, to execute graphical modeling for inferring the gene network.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

Japanese Laid-Open Patent Publication
N . 175305/2002 (*Tokukai* 2002-175305)

A. Relevance of the Above-identified Document

The following is a partial English translation of exemplary portions of non-English language information that may be relevant to the issue of patentability of the claims of the present application.

B. Translation of the Relevant Passages of the Document

See also the attached English Abstract.

[INDUSTRIAL FIELD OF THE INVENTION]

The present invention relates to a graphical modeling method for estimating a gene network, and an apparatus used therefor.

[0002]

[PRIOR ART]

With the recent advance in the DNA micro array technique, it is now possible to simultaneously measure expression of thousands of genes under different conditions, which include different stages of cell cycle, different patterns of cell differentiation, diversity of somatic cells, different clinical conditions, and different species. Some of these different conditions can be ordered. For example, different stages of cell cycle or cell differentiation can be ordered on a time scale.

[0003]

As used herein, the "gene expression profile" is used to refer to the expression level of a group of genes measured under different conditions. By evaluating patterns of expression profile, important information can be obtained concerning gene functions and regulation mechanisms. For example, there has

been developed a method of cluster analysis, in which genes on a microarray are clustered based on similarities of expression profiles. As used herein, the "cluster analysis" refers to the process in which groups of genes that show similar expression patterns on a microarray under different measurement conditions are identified and sorted in clusters. For example, the hierarchical clustering (Eisen et al. 1998), the self-organizing mapping (Tomaya et al. 1999), and other clustering methods (Ben-Dor et al. 1999) have been applied to expression profile data.

[0004]

Gene clustering in expression profiling is useful in predicting functions of gene products, or identifying groups of genes that are regulated by the same mechanism.

[0005]

One of the important information contained in the gene expression profile is the expression network of genes. The expression level of a gene is controlled by other genes, either directly or indirectly. Here, such a network of genes is referred to as a "gene network."

[0006]

Estimating a gene network from expression profiles is one of the important goals of functional genomics. The following methods have been addressed for the estimation of a gene network: Boolean network (Somogyi and Shiegoshi, 1996), a differential equation method (Chen et al., 1999; D'haeseleer et al., 1999), and modeling using a combination of these methods (Akutsu et al., 1998).

[0007]

[PROBLEMS TO BE SOLVED BY THE INVENTION]

Conventional methods include estimation using differential equations in combination with gene disruption

experiments, and modeling based on Boolean modeling. However, neither of these methods has been put to actual application.

[0008]

The present invention was made in view of the foregoing problems, and an object of the present invention is to provide a graphical modeling method that can quickly and accurately estimate a gene network from expression profiles. The invention also provides an apparatus for such a graphical modeling method.

[0009]

[MEANS TO SOLVE THE PROBLEMS]

In order to achieve the foregoing objects, the present invention provides (1) a graphical modeling method for estimating a gene network, the method including the steps of: entering a gene expression profile; calculating a gene correlation coefficient matrix; performing a cluster analysis; outputting a result of cluster analysis; performing calculations for a correlation coefficient matrix of representative genes; converting the correlation coefficient matrix into a partial correlation coefficient matrix; searching a minimum element; re-converting the partial correlation coefficient matrix into a correlation coefficient matrix; determining a significance; and returning to step (f) if there is no significance, and outputting the partial correlation coefficient matrix if there is a significance.

[0010]

In order to achieve the foregoing objects, the present invention provides (2) a graphical modeling apparatus for estimating a gene network, the apparatus including: an input section for entering a gene expression profile; a calculation section for calculating a gene correlation coefficient matrix; a cluster analyzing section; a calculation section for calculating a

correlation coefficient matrix of representative genes; a converting section for converting the correlation coefficient matrix into a partial correlation coefficient matrix; a searching section for searching a minimum element; a re-converting section for converting the partial correlation coefficient matrix into a correlation coefficient matrix; a determining section for determining a significance; and an output section for outputting the partial correlation coefficient matrix, the converting section being operated if there is no significance, and the output section outputting the partial correlation coefficient matrix if there is a significance.

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号
特開2002-175305
(P2002-175305A)

(43)公開日 平成14年6月21日(2002.6.21)

(51)Int.Cl. ⁷	識別記号	F I	テーマコード(参考)
G 0 6 F 17/30	1 7 0 2 1 0	G 0 6 F 17/30	1 7 0 F 4 B 0 2 4 2 1 0 D 4 B 0 2 9
C 1 2 M 1/00		C 1 2 M 1/00	A 4 B 0 6 3
C 1 2 N 15/09		C 1 2 Q 1/68	Z 5 B 0 7 5
C 1 2 Q 1/68		C 1 2 N 15/00	A
審査請求 未請求 請求項の数2 O L (全 11 頁)			

(21)出願番号 特願2000-370947(P2000-370947)

(22)出願日 平成12年12月6日(2000.12.6)

(71)出願人 596013626

株式会社生物分子工学研究所
大阪府吹田市古江台6丁目2番3号

(72)発明者 藤 博幸

大阪府吹田市津雲台3-1 A19-305

(72)発明者 堀本 勝久

佐賀県佐賀市八戸溝3-10-625

(74)代理人 100089635

弁理士 清水 守 (外1名)

Fターム(参考) 4B024 AA11 AA20 HA11 HA19

4B029 AA23 BB20

4B063 QA01 QA18 QQ42 QR55 QR84

QS34 QS36 QS39

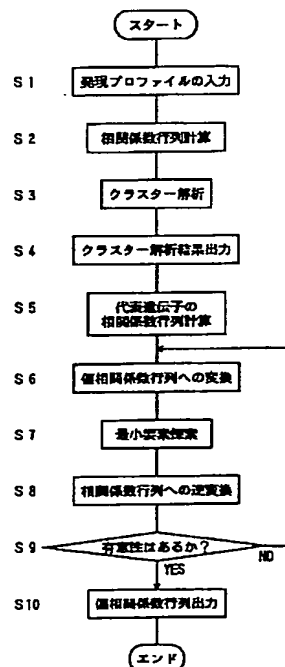
5B075 ND20 NR12 PR06 UU28

(54)【発明の名称】 遺伝子ネットワークを推測するためのグラフィカルモデリング法及びそのための装置

(57)【要約】

【課題】 発現プロファイルから遺伝子ネットワークを推測する遺伝子ネットワークを迅速、かつ的確に推測することができるグラフィカルモデリング法及びそのための装置を提供する。

【解決手段】 遺伝子の発現プロファイルを入力し、遺伝子の相関係数行列の計算を行い、クラスター解析を行い、クラスター解析結果を出力し、代表遺伝子の相関係数行列の計算を行い、偏相関係数行列への変換を行い、最小要素の探索を行い、相関係数行列への逆変換を行い、有意性の判定を行い、有意性の判定の結果、有意性無しの場合には前記偏相関係数行列への変換へ戻り、有意性有りの場合には偏相関係数行列の出力を行い、遺伝子ネットワークを推測するためのグラフィカルモデリングを行う。



【特許請求の範囲】

【請求項1】 遺伝子ネットワークを推測するためのグラフィカルモデリング法において、(a) 遺伝子の発現プロファイルを入力する過程と、(b) 遺伝子の相関係数行列の計算を行う過程と、(c) クラスター解析を行う過程と、(d) クラスター解析結果を出力する過程と、(e) 代表遺伝子の相関係数行列の計算を行う過程と、(f) 偏相関係数行列への変換を行う過程と、

(g) 最小要素の探索を行う過程と、(h) 相関係数行列への逆変換を行う過程と、(i) 有意性の判定を行う過程と、(j) 有意性の判定の結果、有意性無しの場合には前記過程(f)へ戻り、有意性有りの場合には偏相関係数行列の出力を行う過程とを施すことを特徴とするグラフィカルモデリング法。

【請求項2】 遺伝子ネットワークを推測するためのグラフィカルモデリング装置において、(a) 遺伝子の発現プロファイルの入力部と、(b) 遺伝子の相関係数行列の計算部と、(c) クラスター解析部と、(d) 代表遺伝子の相関係数行列の計算部と、(e) 偏相関係数行列への変換部と、(f) 最小要素の探索部と、(g) 相関係数行列への逆変換部と、(h) 有意性の判定部と、(i) 偏相関係数行列の出力部とを備え、(j) 前記有意性の判定部による有意性の判定の結果、有意性無しの場合には前記変換部(e)へ戻り、有意性有りの場合には前記偏相関係数行列の出力部において出力することを特徴とするグラフィカルモデリング装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、遺伝子ネットワークを推測するためのグラフィカルモデリング法及びそのための装置に関するものである。

【0002】

【従来の技術】最近のDNAマイクロアレイ技術の進歩により、異なった状況下で数千の遺伝子の発現を同時に測定することが可能になった。ここでいう異なった状況とは細胞周期の各段階、細胞分化の違い、体細胞の多様性、異なった臨床上で条件あるいは異なった種を指す。この中の幾つかの状況には、順番をつけることが可能である。例えば、細胞周期や細胞分化の段階は時間経過に関して順番つけることができる。

【0003】様々な状況下で測定される一群の遺伝子の発現レベルを、ここでは遺伝子の発現プロファイルと呼ぶ。この発現プロファイルのパターンの評価により遺伝子機能や調節機構に関する重要な知見が得られる。例えば、幾つかのグループは、マイクロアレイ上の遺伝子について、その発現プロファイルの類似性に基づきクラスター分析を行う方法を開発してきている。ここでいうクラスター分析とは、マイクロアレイ上の遺伝子が、異なる測定状況にわたって類似した発現パターンを示す遺伝子群を同定し、クラスターとして分類することをいう。

例えば、階層的クラスタリング(Eisen et al. 1998)や自己組織化マッピング(Tomaya et al. 1999)や他のクラスタリング方法(Ben-Dor et al. 1999)が発現プロファイルデータに応用されてきている。

【0004】発現プロファイルによる遺伝子のクラスター化は、機能が未知の遺伝子産物の機能予測や、同じ機構で調節されている遺伝子群を同定するのに役立つ。

【0005】遺伝子発現プロファイルの中に含まれる重要な情報の一つに、遺伝子間の発現のネットワークがある。一つの遺伝子の発現レベルは、直接的あるいは間接的に他の遺伝子に調節されている。ここではこの様な遺伝子間のネットワークを「遺伝子ネットワーク」と呼ぶ。

【0006】ある発現プロファイルから遺伝子ネットワークを推測することは機能ゲノム学の重要な課題である。Booleanネットワーク(Somogyi and Shiegoski 1996)や微分方程式(Chen et al., 1999; D'haeseleer et al. 1999)やこれらの方法の組み合わせを用いたモデリング(Akutsu et al. 1998)が発現プロファイルから遺伝子ネットワークを推測することは機能ゲノム学の重要な課題である。

【0007】

【発明が解決しようとする課題】しかしながら、従来の方法では、遺伝子破壊実験との組合せによる微分方程式による推定や、ブーリアンモデルによるモデル化が行われていたが、実用の段階にはなかった。

【0008】本発明は、上記状況に鑑みて、発現プロファイルから遺伝子ネットワークを推測する遺伝子ネットワークを迅速、かつ的確に推測することができるグラフィカルモデリング法及びそのための装置を提供することを目的とするものである。

【0009】

【課題を解決するための手段】本発明は、上記目的を達成するために、

〔1〕遺伝子ネットワークを推測するためのグラフィカルモデリング法において、遺伝子の発現プロファイルを入力する過程と、遺伝子の相関係数行列の計算を行う過程と、クラスター解析を行う過程と、クラスター解析結果を出力する過程と、代表遺伝子の相関係数行列の計算を行う過程と、偏相関係数行列への変換を行う過程と、最小要素の探索を行う過程と、相関係数行列への逆変換を行う過程と、有意性の判定を行う過程と、有意性の判定の結果、有意性無しの場合には前記過程(f)へ戻り、有意性有りの場合には偏相関係数行列の出力を行う過程とを施すことを特徴とする。

【0010】〔2〕遺伝子ネットワークを推測するためのグラフィカルモデリング装置において、遺伝子の発現プロファイルの入力部と、遺伝子の相関係数行列の計算

部と、クラスター解析部と、代表遺伝子の相関係数行列の計算部と、偏相関係数行列への変換部と、最小要素の探索部と、相関係数行列への逆変換部と、有意性の判定部と、偏相関係数行列の出力部とを備え、前記有意性の判定部の有意性の判定の結果、有意性無しの場合には前記変換部(e)へ戻り、有意性有りの場合には前記偏相関係数行列の出力部において出力することを特徴とする。

【0011】

【発明の実施の形態】以下、本発明の実施の形態について詳細に説明する。

【0012】本発明では、特に、発現プロファイルから遺伝子ネットワークを推測する新しい方法を示す。これは「グラフィカルモデリング」と呼ばれている方法を応用したものである。グラフィカルモデリングとは多変量解析のための技術の一つで、グラフによって表現された統計的なモデルを推測あるいは検定するために用いられている。本発明では、この「グラフィカルモデリング」の特徴を活かし、またクラスター分析との組み合わせで新規の遺伝子ネットワーク推定の方法を開発した。この方法では、従来の微分方程式を利用した方法のように遺伝子破壊実験などの組み合わせを必要としないという特徴を持つ。つまり、本発明は、グラフィカルモデリングを発現プロファイルに応用した初めての提案である。

【0013】遺伝子ネットワークを推測するための方法は、クラスター分析とグラフィカルモデリングの2段階よりなる。まず解析のために、発現プロファイルデータは遺伝子の相関係数行列に変換される。グラフィカルモデリングは共分散選択(Dempster 1972)を含んでいる。共分散選択には、相関係数行列の逆行列の計算を要求される。発現プロファイルデータ(Eisen et al. 1998)のクラスター分析に示されるように、多くの遺伝子は類似した発現パターンを示す。発現パターンの類似性のため、相関係数行列のいくつかの列または行の間に疑似的な線形従属の関係が生じる。このために逆行列を数値計算によって求めることが困難になる。

【0014】この問題を避けるために、まず、クラスター解析を行い、異なるクラスターに属す遺伝子対の間には疑似的な線形従属の関係が生じないようにクラスターの同定を行う。各クラスターから、代表遺伝子の一つずつ選択した。次に選択された遺伝子の発現プロファイルデータから、相関係数行列を求め、その相関係数行列を用いてグラフィカルモデリングを行った。

【0015】したがって、本発明では、遺伝子間のネットワークではなく、いくつかの遺伝子群の間に形成されるネットワークに関するものである。

【0016】(発現プロファイルデータ)マイクロアレイ上に、L個の遺伝子があるものとする。また、発現レベルの計測が行われた状況の数をNとすると、遺伝子の

発現プロファイルは、 $L \times N$ のサイズを持つ2次元行列Eで表現される。この行列Eの要素(i, j)は、i番目の遺伝子のj番目の状況における発現レベルを示す。発現プロファイルデータはSacchariomyces cerevisiaeの細胞周期にわたるデータをあるウェブサイト(<http://rana.stanford.edu/clustering>)からとって使用した。

【0017】(クラスター分析)ここではクラスター分析のため、堀本らの方法を用いた。ただし、クラスター分析はこの方法に制限されるものではない。すなわち、クラスター分析の結果、各クラスターの構成遺伝子の間で疑似的な線形従属性がなくなるような結果が得られるような方法であれば、別の方法によるクラスター分析の結果も利用できる。

【0018】各クラスターより任意の遺伝子を一個選択し、そのクラスターの代表遺伝子とする。ここでは、L個の遺伝子に1からLの番号をふり、各クラスター中で一番小さな番号を有する遺伝子を代表として選択した。

【0019】(Graphical Gaussian modeling)グラフィカルモデリングを理解するためには、条件付きの独立性の概念が重要である。まずこの概念の説明を行う。

【0020】例として、M個の遺伝子の集合を考える。このM個の遺伝子の発現レベルをある条件下で測定する。ある条件下で計測されたM個の遺伝子の発現レベルはM次元ベクトル $[e_1(\text{Gene } 2), e_1(\text{Gene } 3), \dots, e_1(\text{Gene } M)]$ として表現される。ここでベクトルの要素 e_1 (遺伝子i)は、遺伝子iの発現レベルを示す。fはそのM次元ベクトルの分布の同時密度関数とする。

【0021】発現プロファイルに関して遺伝子AとBが相関を示す場合を考えてみる。2つの遺伝子間で相関が生じるための機構として、次の3つの機構が考えられる。

【0022】第1は、遺伝子間に直接的な相互作用がある場合である。このタイプの相互作用が、遺伝子ネットワークを推定するために必要となるものである。

【0023】第2は、2つの遺伝子の間に間接的な相互作用がある場合である。言い換えると、遺伝子Aの産物の調節の情報、ほかのいくつかの遺伝子の発現を介して、遺伝子Bの発現を引き起こすように伝達されている場合である。

【0024】第3は、共通の遺伝子により調節されているために生じる相関である。

【0025】遺伝子ネットワークを推測するためには、第一のタイプの相互作用とはかのタイプの相互作用を区別することが重要である。

【0026】次に、M次元ベクトルの同時密度関数を考える。この同時密度関数が、一つは e_1 (遺伝子B)を

10

20

30

40

50

含まず、他方は $e1$ (遺伝子A)を含んでいないような二つの要素に因数分解される時、 $e1$ (遺伝子A)と $e1$ (遺伝子B)はそれ以外のものが与えられたときに条件付き独立であるという。

【0027】式: $f[e1(Gene1), e1(Gene2), \dots, e1(GeneA), \dots, e1(GeneB), \dots, e1(GeneM)] = h[e1(GeneA), the\ rest] \times k[e1(GeneB), the\ rest]$,

ここで、 h は $e1$ (遺伝子B)を含まない関数、 k は $e1$ (遺伝子A)を含まない関数である。このルールは因数分解基準(Dawid, 1979)とよばれている。条件付き独立性は、2つの発現レベルが直接には関係していないことを意味している。これ以降、記述の煩雑さを避けるため、 $e1$ (遺伝子A)と $e1$ (遺伝子B)の間に条件付き独立が成立することを、遺伝子Aと遺伝子Bは条件付き独立であると呼ぶことにする。二つの遺伝子が条件付き独立である時には、それらの間に相関があっても直接の相互作用はないことを意味している。ここで、 M 個の遺伝子の間の条件付き独立関係を表現するために、グラフ $g = (v, e)$ を考える。 v は、 M 個の遺伝子に対応する頂点の有限集合であり、 e は2つの頂点間を結ぶ辺の有限集合である。

【0028】グラフ g は、その発現レベルが条件付き独立であるような遺伝子に対応する頂点間の辺を除去するように因数分解基準を繰り返し適用することによって構築することができる。このとき、 e は他の要素が与えられた時に条件付き独立でないような発現レベルを有する遺伝子ペア間を結ぶ辺により構成されている。得られたグラフは、 M 個の遺伝子の間の遺伝子ネットワークを表現すると見なすことができる。ここでは、先に述べたクラスター分析により、数が減じられた遺伝子の発現プロファイルデータにグラフィカルモデリングを適用することを考える。それらの遺伝子の発現プロファイルデータは、 $M \times N$ の2次元実行列で表される。ここで、 M はクラスター分析によって減じられた遺伝子の数であり、 N は、その発現レベルが計測された条件の数である。いうまでもなく、 $L > M$ である。モデルを構築するために、 M 個の計測値 $e1(Gene1)$ から M は以下の同時分布関数を持つランダムなベクトルであると仮定する。

【0029】式: $f\theta[e1(Gene1), e1(Gene2), \dots, e1(GeneA), \dots, e1(GeneB), \dots, e1(GeneM)]$,

ここで、 θ は、未知のパラメータである。その時に、先に述べた発現プロファイルを表す2次元行列は、 $f\theta$ からの独立な N 回の観測から得られた標本とみなすことができる。また、パラメータ θ は、遺伝子の発現レベルの間の条件付き独立性についての情報を含むようにデザインされているものとする。それゆえに、グラフの構造の推定は、パラメータ θ の推測を通じて行われる。しかし

ながら、 f や θ による密度関数の記述はあまりにも一般的すぎて、具体的な推定を行うことができない。連続データに関するグラフ構造の推定のためには、 f や θ として多変量正規分布が仮定される。この時 θ は、分散行列 Σ と平均ベクトル μ より構成されている。多変量正規分布に基づいた方法は、グラフィカル・ガウシアン・モデリングと呼ばれている。ここでは、 M 個の遺伝子の発現レベルは多変量正規分布に従うランダムベクトル $[e1(Gene1), e1(Gene2), \dots, e1(GeneM)]$ であると仮定する。

(独立グラフの推定) 多変量正規分布の仮定の下では、因数分解基準はある二変数の間で偏相関係数が0になるということと等価である。すなわち、2つの変数はその偏相関係数が0であるときに限り、条件付き独立となる。偏相関係数を得るためには、共分散行列の逆行列の計算が要求される。

【0030】($\Omega = \Sigma^{-1}$) $\Omega = (\omega_{ij})$ が与えられたときに、 $e1(Genei)$ と $e1(Genej)$ の間の偏相関係数 $\rho^{ij, the\ rest}$ は、 Ω の要素から以下のように計算される。 $\rho^{ij, the\ rest} = -\omega_{ij} / (\omega^{ii} \times \omega^{jj})$ として与えられる。この式が示しているように、偏相関係数が0であることは、逆共分散が0であることに対応している。グラフィカル・ガウシアン・モデルは、特定の偏相関係数を0にするようなモデルとして定義される。遺伝子間の条件付き独立の関係を評価することによって、グラフ構造を構築するために、ここでは、WermuthとScheidtらによって開発された、段階的な繰り返しアルゴリズムを用いた。共分散行列や逆共分散行列の代わりに、相関係数行列と偏相関係数行列を用いた。

【0031】ステップ0: 初期化のステップである。完全グラフ $g(0) = (v, e)$ を用意する。このグラフの頂点は M 個の遺伝子に対応しており、全ての頂点は連結されているものとする。この $g(0)$ をフルモデルと呼ぶ。発現プロファイルデータに基づいて最初の相関係数行列 $C(0)$ を計算する。ただし、 $\tau = 0$ とする。

【0032】ステップ1: 相関係数行列 $C(\tau)$ から偏相関係数行列 $P(\tau)$ を計算する。ここで、 τ は繰り返しの回数を示す。

【0033】ステップ2: $P(\tau)$ の全ての要素の中からその絶対値が最小である要素を探して、その要素を0と置き換える。

【0034】ステップ3: $P(\tau)$ から相関係数行列 $C(\tau+1)$ を再構築する。 $C(\tau+1)$ においては、 $P(\tau)$ において0に置き換えられた要素に対応する要素のみが変更される。一方、他のすべての要素は、 $C(\tau)$ の要素と同じである。

【0035】ステップ4: 繰り返し計算の打ち切りの判断を行う。そのために以下の3つの数値を計算する。

【0036】

$$\begin{aligned} \text{dev1} &= N \log [|C(\tau+1)| / |C(0)|] \\ \text{dev2} &= N \log [|C(\tau+1)| / |C(\tau)|] \end{aligned}$$

$|C(\tau+1)|$, $|C(\tau)|$, $|C(0)|$ は、各行列の行列式を表す。 dev1 は自由度= n における χ^2 分布に従い、 dev2 は自由度= 1 における χ^2 分布に従う。 n は τ 回の反復まで0にセットされた要素の数を示す。 N は計測が行われた状況の数を表す。

【0037】ステップ5： dev1 あるいは dev2 以上の値が生じる確率を、それぞれ χ^2 分布に従い求めた時、 dev1 に対する確立値が0.5以下か、 dev2 に対する確立値が0.3以下であれば、 $C(\tau+1)$ に対応するモデルは棄却され、繰り返し計算を終了する。終了条件が満たされなかった時は、 $P(\tau)$ において偏相関係数が0にセットされた2つの遺伝子間の辺は $g(\tau)$ から除去して、 $g(\tau+1)$ とし、 τ を $\tau+1$ に進めて、ステップ1に戻る。

【0038】上記の方法で最終的に得られたグラフは無向グラフでランダム変数間の条件付き独立関係を表現している。この無向グラフを独立グラフと呼ぶ。

【0039】以下、本発明の実施例について説明する。

【0040】図1は本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング装置の構成図である。

【0041】この図において、1は入力装置、2は発現プロファイル入力部、3は相関係数行列計算部、4はクラスター解析部、5は代表遺伝子の相関係数行列計算部、6はグラフィカル・ガウシアン・モデリング実行部、7は偏相関係数行列への変換部、8は最小要素探索部、9は相関係数行列への逆変換部、10は有意性判定部、11は偏相関係数行列出力部、12は出力装置、13はCPU（中央処理装置）、14は記憶装置である。

【0042】図2は本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング法のフローチャートである。

【0043】本発明のグラフィカルモデリング法の概要について図2を参照しながら説明する。

【0044】①遺伝子の発現プロファイルを入力し（ステップS1）、次に、遺伝子の相関係数行列の計算を行い（ステップS2）、次に、クラスター解析を行い（ステップS3）、次に、クラスター解析結果を出力し（ステップS4）、次に、代表遺伝子の相関係数行列の計算を行い（ステップS5）、次に、偏相関係数行列への変換を行い（ステップS6）、次に、最小要素の探索を行い（ステップS7）、次に、相関係数行列への逆変換を行い（ステップS8）、次に、有意性の判定を行い（ステップS9）、有意性の判定の結果、有意性無しの場合には前記ステップS6へ戻り、有意性有りの場合には偏相関係数行列の出力を行う（ステップS10）。

【0045】以下、具体例を詳細に説明する。

【0046】（1）クラスター分析

酵母の2,467個の遺伝子を、その発現プロファイルに関する相関係数行列の各行あるいは各列が、クラスター間では数値解析の意味において一次独立になるようにクラスターを構成した。

【0047】その結果、2,467個の遺伝子は34個のクラスターに分類された。

【0048】（2）グラフィカル・ガウシアン・モデリング（Graphical Gaussian Modeling, GGM）

上記の34クラスターの代表遺伝子の発現プロファイルの相関係数行列を用いてGGMを実行した。

【0049】GGMの繰り返し計算の過程における、 dev1 と dev2 のp-valueの変化を繰り返しのステップに対しプロットしたものを、図3に示す。すなわち、図3では、 dev1 と dev2 の確立値の繰り返しのステップ番号に対してのプロット dev1 と dev2 の確立値は、GGMの繰り返し計算の停止の判定のために使用されており、それを繰り返しのステップ数の関数として、プロットしている。●は dev1 の確立値、○は dev2 の確立値を表す。

【0050】この図から明らかなように、 dev1 のp-valueは、図中aで示すように、全過程において99.99%より大きく、繰り返しの打ち切り評価には利用されなかった。一方、 dev2 のp-valueは、図中bで示すように、繰り返しステップが進むにつれて減少してゆき、137ステップ目で偏相関係数行列（PCCM）の要素（9,28）の評価の際、p-valueが0.284となり設定した基準値0.3より小さくなったため繰り返しは打ち切られた。したがって、PCCM中136個の要素が0.0に置換されたことになる。以上のように、GGMの繰り返し計算の打ち切りには、 dev2 の方が効率的であった。

【0051】GGMによって得られた34個の代表遺伝子の最終的なPCCMを図4に示す。すなわち、図4には、GGMによって得られた偏相関係数の行列GGMの繰り返しの過程で0.0に置換された要素は網がけで示されている。

【0052】要素（7,8）がPCCM中最小で、-0.75、要素（4,24）が最大で0.59であった。561個のPCCMの要素の内、136個が0に置換された（約24%）。言い換えると、136個の辺が独立グラフから削除されたことになる。独立グラフには孤立した節はなく、また全ての他の節に辺を有する節もなかった。節の内、最多の辺を有するものでは、30個、最少の辺を有するものでは18個の辺を有していた。独立グラフで全ての辺を表示するのは煩雑になるため、ここでは偏相関係数の絶対値が0.5より大きなもののみ表示したものを図5に示す。すなわち、図5に

は、34個のクラスターの独立グラフが示されており、絶対値が0.5より大きな偏相関係数に対応する辺のみが示されている。

【0053】以下、さらに議論を深めることにする。

【0054】(1) 代表遺伝子のGGMのための妥当性
2, 467個の遺伝子の発現プロファイルデータのサンプル数がさらに増加すれば、より小さなメンバーから構成されたクラスターへの分割が可能になるかもしれない。しかし、ここでは、得られた34クラスターの各々に含まれる遺伝子は、その発現の挙動が類似している、すなわち類似した制御を受けていると仮定し、34個の遺伝子のPCCM(図4)と独立グラフ(図5)は、34個のクラスター間のPCCMと独立グラフを表しているものとして議論をすすめる。

【0055】その前に、代表遺伝子の選択の妥当性を述べておく。代表遺伝子を選択する様々のものが考えられるが、そのような方法の違いが、GGMの結果に影響を与えないか否かについて検討する。

【0056】まず、先に選択した代表遺伝子以外の34クラスターのメンバーをランダムに選択して、100個の34個の代表遺伝子のセットを用意した。そして、その各々と、先に選択した代表遺伝子について、その発現プロファイルの相関係数行列の行列式の比(比が1より小さい場合はその逆数)の対数に561を乗じたものは自由度561のカイ自乗分布に従うと予想されるので、そのp-valueを計算してみた。有意水準を0.05とすると、100個中7個のセットのみがオリジナルの代表遺伝子セットとは有意に異なるものとして棄却された。また、GGMでdev1、dev2による繰り返しステップの打ち切りを使用している有意水準、0.3と0.5を使用しても、100セットの内9セットと11セットがそれぞれオリジナルの代表遺伝子とは有意に異なるものとして棄却された。この結果は、必ずしも異なる代表遺伝子セットを用いたGGMの結果が一致することを保証はしないものの、一組の代表遺伝子によってGGMを行った結果で十分であることを強く示唆している。

【0057】(2) 独立グラフによる辺の解釈

上記したように、ここは、PCCMの非0要素あるいはそれに対応した独立グラフの辺は、クラスター間の直接的な相互作用を表しているものと仮定している。このセクションでは、辺の解釈について吟味する。一般的にPCCMの非0要素は変数のペアの間の直接的な相互作用を示唆している。しかし、それはどちらの変数が原因でどちらが結果であるかは示すことができない。それ故に、独立グラフ中の辺は無向である。因果関係に対する先験的な知識無しには辺を矢印に置き換えることはできない。図4に示されているPCCMの非0要素はクラスターのペアの間の直接的な相互作用を表すものと考えられている。そのときに、クラスター間の直接的な相互作用

用がなにを意味するかについて説明する。

【0058】あるクラスターに含まれているすべての遺伝子が、辺で結ばれている他のクラスター内のすべての遺伝子の発現に影響するということは現実的であるとは思えない。むしろ、あるクラスターの遺伝子の部分集合が他のクラスターの遺伝子の部分集合の発現に影響しているということの方が考えやすい。更に、クラスター間の辺はループを構成しているかもしれない。例として、1つの辺で結ばれたクラスターA、Bで考えてみよう。クラスターAの遺伝子の部分集合にクラスターBの遺伝子の部分集合が発現に影響を及ぼしており、一方、クラスターBの他の遺伝子の部分集合がクラスターAの部分集合に影響を与えていることが考えられている。このように、クラスター間の辺が遺伝子間に及ぼす影響は、直接的な相互作用よりさらに複雑な意味を持っているかもしれない。加えて、ここで検討された発現プロファイルデータには、機能について未知の多くの遺伝子が含まれていないということを心に留めておかなければならないだろう。

【0059】アイゼンらの仕事(1998)においては、その機能が同定されている酵母の2467個の遺伝子の発現レベルが計測されている。しかし、ルビンら(2000)によれば、酵母のゲノム中には6241個の遺伝子が蛋白をコードされるものとして存在することが予想されている。このように、データが欠けていることによって、いくつかの辺は直接的な相互作用ではなく、機能未同定の遺伝子の発現を通じて、間接的に相互作用している遺伝子のために生じたという可能性を除去することができない。

【0060】(3) 推定された遺伝子ネットワークの特徴

独立グラフは孤立点を含まなかった。また、どの節も他のすべての節に対して、辺を持つものはなかった。遺伝子ネットワークについての重要な問題の1つは、遺伝子ネットワークはいくつかの独立した部分ネットワークに分割できるかどうかということである。この問題は、独立グラフが他の部分グラフの節に対して辺を持たないようないくつかのサブグラフに分割できるかどうかということで、容易に確認することができる。得られたPCCM(図4)の検討から、そのような分割はできないということが示唆された。すなわち、遺伝子ネットワークは単一のシステムとして働いている。推定された遺伝子ネットワークの調節機構を特徴づけるため、PCCMをヒストグラムに要約して示したものが図6である。すなわち、図6では、偏相関係数の頻度分布が示されている。横軸は偏相関係数を表している。頻度の計算のため、偏相関係数は、0.1の幅を持った離散データにまとめられた。縦軸は、幅内の係数の頻度を表している。しかしながら、偏相関係数の0.0の頻度は横軸上の0.0に対してプロットされている。

【0061】この図6のヒストグラムでは、偏相関係数の頻度分布が示されている。PCCMの561個の要素中、239個が正の値を持っており、一方、186個は負の値を持っていた。そして、その数はそれぞれ43%、33%に対応していた。さらに、正の偏相関係数の分布は、負の相関係数の分布に対して、対称ではなかった。正の偏相関係数の分布は、0.2から0.3までの範囲にピークを持っていた。さらに、0.0から0.1の範囲の相関係数は存在しなかった。これに対して、負の偏相関係数の分布のピークは、-0.2から-0.3にあった。分布の端は、正の相関係数の場合に比べ、両側にのびていた。この観察は、正の調節の数の方が負の調節の数よりも多いことを示唆している。

【0062】mRNAの転写に関与している遺伝子は、遺伝子の発現の調節に関わっている可能性が最も高い。それ故にあるクラスター中の転写関連遺伝子の集合は、辺によって結ばれている他のクラスターの遺伝子の発現の調節において中心的な役割を演じているかもしれない。34のクラスター中の、33個はmRNAの転写に関与している可能性を持っていた。しかしながら、その量はクラスターごとに異なっていた。この観察はmRNAの転写が並列分散型の調節を受けているということを示唆している。PCCM（図4）や独立グラフ（図5）は、ある程度まではmRNAの発現に関与している

遺伝子による発現の調節を表しているのかもしれない。一方、クラスター32は、翻訳関連遺伝子の40%以上と、rRNAの転写関連遺伝子の約50%を含んでいた。クラスター32に結ばれている辺の数は28であり、5つの辺がGGMの過程で除かれたことになる。他のクラスターにおいては、翻訳関連遺伝子の含量は低かった。したがって、翻訳はむしろ、集中的な調節系を採用しているのかもしれない。

【0063】ここでは、細胞周期調節遺伝子に注目して、本発明のモデルネットワークを簡略化してみた。シュベルマンらは、294個の細胞周期調節遺伝子を酵母中で同定していた。この研究に使われた2467個の遺伝子のうち、232個の遺伝子がシュベルマンらによって分類されている細胞周期調節遺伝子と対応していた。細胞周期のステージはG1、S、G2、M、M/G1の5つがあり、その各々が周期に特異的に発現している細胞周期因子を有している。232個の遺伝子のうち、93、32、28、48、31個の遺伝子がそれぞれG1、S、G2、M、M/G1に関連していた。ここで、34個のクラスターの中におけるそれらの遺伝子の分布を調べてみたものが表1である。

【0064】

【表1】

13

14

クラスター	G 1	S	G 2	M	M/G 1	合計
1	0	0	0	1	0	1
2	1	5	1	1	1	9
3	0	0	0	0	0	0
4	1	1	1	1	0	4
5	0	0	1	0	0	1
6	0	0	0	0	0	0
7	15	4	10	4	2	35
8	1	1	1	0	0	3
9	8	0	1	5	1	15
10	0	0	0	1	2	3
11	12	9	0	1	2	24
12	36	6	3	2	0	47
13	1	0	0	1	0	2
14	0	1	0	4	1	6
15	2	0	1	1	0	4
16	0	1	0	0	1	2
17	0	0	0	0	0	0
18	0	0	0	0	2	2
19	0	0	0	0	1	1
20	1	1	0	1	3	6
21	1	0	0	0	0	1
22	1	0	0	1	1	3
23	1	0	2	2	4	9
24	1	0	2	2	1	6
25	1	0	1	9	0	11
26	1	1	0	1	2	5
27	1	0	0	2	1	4
28	3	0	0	0	1	4
29	0	1	0	0	0	1
30	1	1	0	4	2	8
31	2	0	0	1	3	6
32	1	0	2	2	0	5
33	0	0	0	0	0	0
34	1	0	2	1	0	4
	93	32	28	48	31	232

【0065】34個のクラスター中、4個は何の細胞周期調節遺伝子も含んでいなかった。クラスター12は、G1に関連している細胞周期関連遺伝子を最も多く含んでいた。同様に、クラスター7と25は、それぞれG2とM2に関連している細胞周期関連遺伝子を最も多く含んでいた。SとM/G4に関連した細胞周期関連遺伝子は、いくつかのクラスターに分散して存在しているように見えるが、比較的クラスター11と12に集中していた。ここではクラスター7、11、12、23、25に関連したPCCMの要素を集めてみた。それを図7に示す。すなわち、図7には、クラスター7、11、12、23、25の間の偏相関係数が示されている。

【0066】図7に示すように、9つの要素のうち、6個は0だった。また、2つの要素(11、25)と(12、25)は、0.11という小さな値を持っていた。クラスター11と12のペアは0.37、クラスター7と23のペアに対応する要素のみが-0.32という大きな値を示していた。この観察は、細胞周期の異なるステージに関連している細胞周期調節遺伝子が基本的には

お互いに条件付き独立であるか、非常に弱い関係しか持たないということを示唆している。

【0067】なお、本発明は上記実施例に限定されるものではなく、本発明の趣旨に基づいて種々の変形が可能であり、これらを本発明の範囲から排除するものではない。

【0068】

【発明の効果】上記したように、本発明によれば、発現プロファイルから遺伝子ネットワークを推測するグラフィカルモデリングにより遺伝子ネットワークを迅速、かつ的確に推測することができる。

【図面の簡単な説明】

【図1】本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング装置の構成図である。

【図2】本発明の実施例を示す遺伝子ネットワークを推測するためのグラフィカルモデリング法のフローチャートである。

【図3】GGMの繰り返し計算の過程における、dev

15

1とdev2のp-valueの変化を繰り返しのステップに対しプロットした図である。

【図4】GGMによって得られた34個の代表遺伝子の最終的なPCCMを示す図である。

【図5】34個のクラスターの独立クラスター。偏相関係数の絶対値が0.5より大きなもののみ表示した図である。0.5～0.55が異なる長さのセグメントよりなる破線。0.55～0.6が同じ長さのセグメントよりなる破線。0.6～0.7は細い実線。0.7～0.8は太い実線で示されている。

【図6】推定された遺伝子ネットワークの調節機構を特徴づけるため、PCCMをヒストグラムの形に要約して示した図である。横軸は偏相関係数を表している。頻度の計算のため、偏相関係数は、0.1の幅を持った離散データにまめられた。縦軸は、幅内の係数の頻度を表している。しかしながら、偏相関係数の0.0の頻度は横軸上の0.0に対してプロットされている。

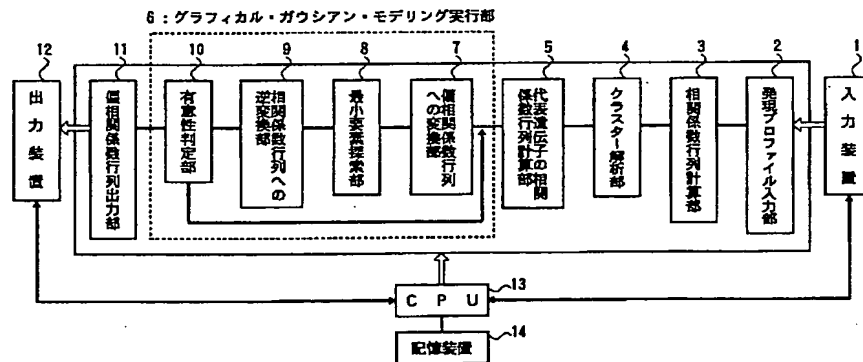
*

*【図7】クラスター7、11、12、23、25に関連したPCCMの要素を集めた図である。

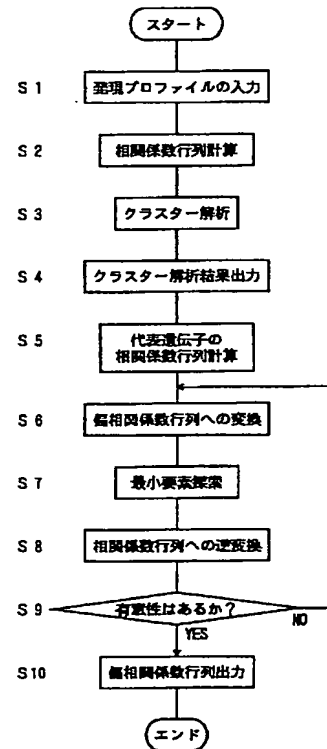
【符号の説明】

- 1 入力装置
- 2 発現プロファイル入力部
- 3 相関係数行列計算部
- 4 クラスター解析部
- 5 代表遺伝子の相関係数行列計算部
- 6 グラフィカル・ガウシアン・モデリング実行部
- 7 偏相関係数行列への変換部
- 8 最小要素探索部
- 9 相関係数行列への逆変換部
- 10 有意性判定部
- 11 偏相関係数行列出力部
- 12 出力装置
- 13 CPU（中央処理装置）
- 14 記憶装置

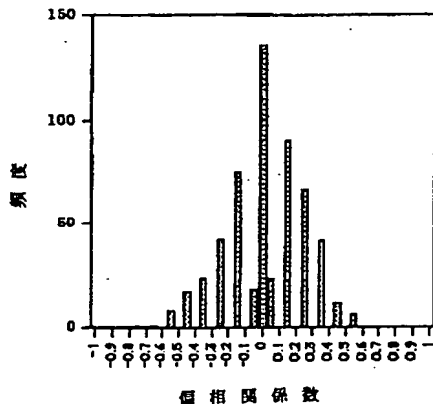
【図1】



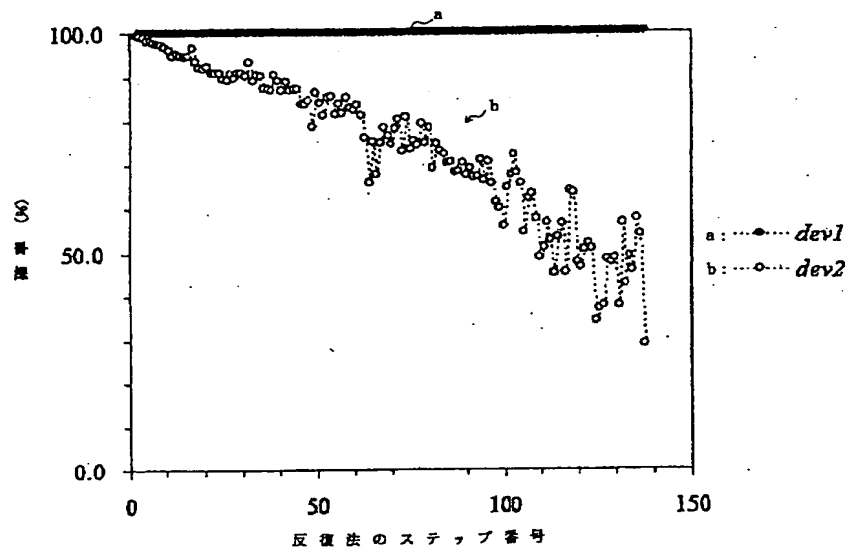
【図2】



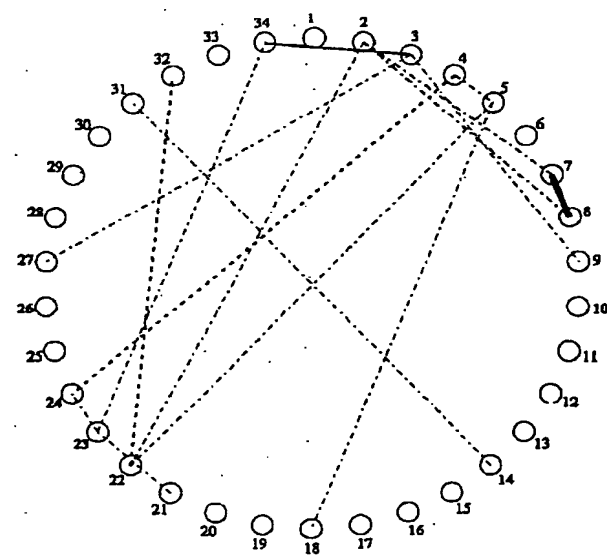
【図6】



【図3】



【図5】



【図7】

G1	12	-				
S	11	0.37	-			
G2	7	0.003	0.003	-		
M	25	0.11	0.11	0.003	-	
M/G1	23	0.003	0.003	-0.32	0.003	-
		12	11	7	25	23

[illegible]